Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 998 – 1002

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

# Delphi Inductive Algorithm Realization Based on the Statistical Theory

Guogang Li[a*], Yunhong Lia[a], Guoqiang Lib[b]

[a]*College of Sciences, Hebei University of Science and Technology, Shijiazhuang 050018, China*
*b Shijiazhuang City Science and Technology Information Institute, 050018, Shijiazhuang, China*

**Abstract**

Machine learning is an important research field of application of artificial intelligence, classification algorithms is an important technology in data mining, the decision tree learning is a common method of this paper, Based on the details of typical decision tree classification algorithm - ID3 algorithm that commonly used in the field of machine learning and data mining,using the information of the statistical law that examples provided to us which simplifies the classified pruning and decision tree optimization process, and realize this algorithm by Delphi simulation experiment, has the accuracy is high, classification speed characteristics. It gives a basic principles, methods, execution, correctness proof and Performance Analysis of Algorithms.



*Keywords: Induce learning; Classificication algorithm; Statistical rule; Data mining; decision tree; BS-CA.*

## 1. Introduction

Since the artificial intelligence became an independent discipline,machine learning has become since the center in the field research topic. Machine learning is a discipline to study how to use the machine to simulation of human activities, his core is study. Relatively strict formulation is: machine learning is a new skills of science research and acquire new knowledge, and it is a knowledge identify the existing knowledge. Knowledge is the key to intelligence system,such as expert system and intelligent decision support system. At present, machine learning method is not practical, knowledge acquisition mainly dependent on knowledge engineers and experts communication. Due to the subjective and objective factors, the knowledge gained from the method is plain get, insufficient, influence the performance of the

* Guogang Li. Tel.:86+015032076885.
*E-mail address*: lgg-2000@163.com.


doi:10.1016/j.proeng.2012.01.078

system. Especially expert empirical knowledge, it is difficult to description use language, but their effect is important. Experts experience gained from the plenty of examples. Therefore, collecting examples, and obtain knowledge directly, is a effective way to solve knowledge acquisition bottles of diameter.Learn from examples, already appeared many methods, such as genetic algorithm[1] , and concept tree decision tree[2-4] and rough set theory, etc.

Data Mining (Data Mining) is a process that extract implicit in advance again, unknown, but potentially useful information and knowledge utilization analysis tools from a lot of, incomplete, noisy, fuzzy and random Data[5-6], and the process find the relationship between from models and data, make prediction. At present the studies of data mining are mainly concentrated in Classification, clustering, association rules mining, sequential patterns found, abnormal and trends found etc[7].

## 2. Algorithmic

Quinlan puts C4.5 algorithm forward in 1993, which also use the information increment as a selection criterion attributes as well. It inherited all the advantages of ID3 algorithm and made some improvements. C4.5 can be divided into two stages: First, according to the criteria of largest information increment choosing identity attribute for the division of training sets, recurring until all the examples in each division belonging to the same category; then, make a pruning on the tree which Newly created, which is cutting the branches based on the noise data[8].

### *2.1. Structure*

C4.5 algorithms have simple idea, efficient algorithms, and reliable results', but it stills have some deficiencies. First, C4.5 takes a divide-and-rule strategy, constructing the internal nodes in the tree by the partly optimal way, so despite of the high accuracy of the final results, still fall short of the overall best results; Secondly, The main basis for the evaluation of decision tree decision tree error rate, and the depth of the tree the nodes are not considered. And the average depth of the tree directly corresponds to the forecast rate of decision tree, tree node number is the size of the tree, following are a few representatives of the node tree size; thirdly, as constructing decision tree[9], as evaluating it, it is difficult to adjust the structure and content of the tree, and it is very difficult to improve its performance, C4.5 attribute value groups one by one, without heuristic search mechanism, the efficiency is low[10].

### *2.2. The structure of SD-CA*

Set A is a attribute sets, each sample has k attributes, A = {A1, A2,... AK}, among them, attribute A1 has l1 attribute values, the attribute A2 has l2 attribute value... Attribute Ak has lk a attribute value, namely

| attribute | Attribute value |
|---|---|
| A1 | A11,A12,…,A1l1 |
| A2 | A21,A22,…,A2l2 |
| … | … |
| Ak | Ak1,Ak2,…,Aklk |

P is positive examples class in Sample set, N is negative patients with type in Sample se.

the detailed steps of the Algorithm are as follows:

(1) hypothesis of return for positive cases, the calculation:

$$\mu(A_{ij},P)=\frac{P(A_{ij})}{A_{ij}}$$

$$\xi(A_{ij},P)=\frac{P(A_{ij})}{P(A_i)}$$

Hypothesis of return for positive cases, the calculation:

$$\mu(A_{ij},N)=\frac{N(A_{ij})}{A_{ij}}$$

$$\xi(A_{ij},N)=\frac{N(A_{ij})}{N(A_i)}$$

$$(\mathrm{i}=1,2,\cdots,l_i;\ \mathrm{j}=1,2,\cdots,k)$$

（1）

$$\mu\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\right)=\mu\left(A_{1s_1},P\right)\mu\left(A_{2s_2},P\right)\cdots\mu\left(A_{ks_k},P\right);\ \xi\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\right)=\xi\left(A_{1s_1},P\right)\xi\left(A_{2s_2},P\right)\cdots\xi\left(A_{ks_k},P\right);$$

$$\mu\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},N\right)=\mu\left(A_{1s_1},N\right)\mu\left(A_{2s_2},N\right)\cdots\mu\left(A_{ks_k},N\right);\ \xi\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},N\right)=\xi\left(A_{1s_1},N\right)\xi\left(A_{2s_2},N\right)\cdots\xi\left(A_{ks_k},N\right)$$

$$\mathrm{s}_1=1,2,\cdots,l_i;\ \mathrm{i}=1,2,\cdots,k$$

（2） Compare:

$$\mu\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\right)\xi\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\right)$$

$$>\mu\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},N\right)\xi\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},N\right),\text{then } A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\in P;$$

$$\mu\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\right)\xi\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\right)$$

$$<\mu\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},N\right)\xi\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},N\right),\ A_{1s_1},A_{2s_2},\cdots A_{ks_k},N\in N;$$

（3） Confidence

$$\beta\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\right)$$

$$=\frac{\mu\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\right)\xi\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\right)}{\mu\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\right)\xi\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},P\right)+\mu\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},N\right)\xi\left(A_{1s_1},A_{2s_2},\cdots A_{ks_k},N\right)}$$

## 3. Performance Analysis

### *3.1. Index Analysis*

The below table 1 gives the data collections of influence the comfort of summer weather for several related index, it has four attributes: clad index, temperature, humidity, wind. These four attributes are divided into comfortable (positive cases) and discomfort (counterexample) two kinds. The following will be the sample data set for the training set, we class the data use ID3 algorithm and the algorithm of this article.

Table.1 Training set of sample

| attribute | clothing | temperature | humidity | wind | Comfort levels |
|---|---|---|---|---|---|
| 1 | much | higher | bigger | none | N |
| 2 | much | higher | bigger | big | N |
| 3 | much | higher | bigger | middle | N |

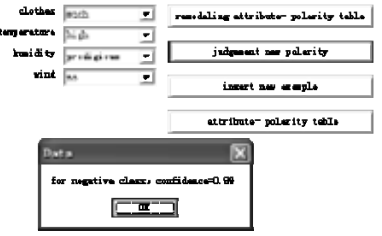
| | | | | | |
|---|---|---|---|---|---|
| 4 | normal | higher | bigger | none | P |
| 5 | normal | higher | bigger | middle | P |
| 6 | many | moderate | bigger | none | N |
| 7 | many | moderate | bigger | middle | N |

To classify the Training set of sample use ID3 algorithm, get the decision tree,we can see from Table 3, The first 4 sample have the same classification results , and The latter two is different; According to the rules of ID3, when the 5th example dressing Index is high, classification is only related to emperature,therefore, it is positive examples when the temperature is high; when the 6th example dressing Index is high, classification is only related to humidity, therefore, it is positive cases of normal humidity. And this algorithm combined different proportional value of the four property values, and obtain the classification results. Through the confidence,determine of the accuracy of one of our relative classification.

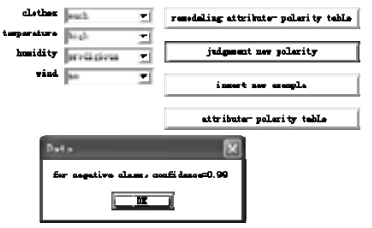
According to the decision tree, for the path from the root node to leaf nodes and the quantity of the record in data set, can be drawn as follows classification rules, Form says such as table 2.

(1) (IF (clad = more AND humidity = normal) THEN (category =positive)
(2) (IF (clad = more AND humidity = great) THEN (category = negative)
(3) (IF (temperature = normal AND clad = more) THEN (category = negative)
(4) (IF (clad = more AND temperature = high wind = great) THEN (category = negative)

The following algorithm using the sample data set of Table 3.1 for training test with Delphi




（figure 1.）




（figure 2.）

We can see from the above image, different attribute value on classification play different roles, some attribute values are cases in class and counterexample class the equal proportion , for example, when the wind for medium , ratio were 0.5, then scale to all, its classification results more depends on the other attributes of the attribute value selection; Some attribute value have decisive role on sample classification , for example, Dressing index as long as "normal" value of this property, Then the cases will be positive. Of course, more attribute value plays a different weight role in classification, Therefore, comprehensive consideration, the final results obtained. By probability statistics knowledge, The proportion of positive and negative property values becomes stable， the accuracy is higher in the classification.

### *3.2. testing analysis*

Here is the comparison between selected sample set of tests and testing results of ID3.

Table 2. test results comparison

| | clothing | temperature | humidity | wind | class | confidence | ID3 |
|---|---|---|---|---|---|---|---|
| 1 | many | higher | bigger | bigger | N | 0.99 | N |
| 2 | many | moderate | normal | bigger | N | 0.96 | N |
| 3 | normal | higher | bigger | bigger | P | 0.94 | P |
| 4 | much | moderate | bigger | bigger | N | 0.99 | N |
| 5 | many | higher | bigger | no | N | 0.99 | P |
| 6 | much | higher | normal | bigger | N | 0.89 | P |

we can see From table 2, The first 4 sample have the same classification results , and The latter two is different; According to the rules of ID3, when the 5th example dressing Index is high, classification is only related to temperature,therefore, it is positive examples when the temperature is high; when the 6th example dressing Index is high, classification is only related to humidity, therefore, it is positive cases of normal humidity.. But the SD - CA algorithm combined different proportional value of the four property values , and obtain the classification results. Through the confidence, determine of the accuracy of one of our relative classification.

## 4. Conclusions

Through the experimental analysis, and compared the ID3 algorithm, this paper SD - CA algorithm and ID3 generate decision tree algorithm has very big different, get classification rules also vary, some also with actual abhorrent, this and data contained noise and related data sets and too small. But the algorithm to a certain extent of weather reduced clad index from the importance of measurable impact root node (long distance), the corresponding increased humidity, temperature, wind attribute the importance in classification, avoid the decision tree algorithm may be converging local optimal solution and lost the global optimal solution shortcomings. Meanwhile, this algorithm avoids the ID3 algorithm in more tedious logarithm process simple calculation, and it directly attribute value and are examples and the negative cases of link, by comparing the results of a certain attribute value for a positive or negative for classification of probability, the corresponding determine its classification results and are able to comment directly, thus a simple calculation process, reducing the computational complexity, improve the calculation accuracy.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China (71071049, 70871036) and the Natural Science Foundation of Hebei Province (F2011208056).

## References

[1]Chen Wenwei. Support System of Intelligent Decision. *Beijing: Electronic Industry Press*,1998,pp.89-101.

[2]Quinlan J. R. Induction of decision trees. *Machine Learning,* 1986, pp.81-106.

[3]Hong J. R. A El: an extension approximate method for general covering problem .*International Journal of Computer and Information Science, 1985*, pp.421-457.

[4]Brodley C. E. , U tgoff P. E. Multivariate decision trees. *Machine L earning,* 1995, pp.45-77.

[5]Zeng Huanglin,Rough Set Theory and Its Applications. *Chongqing: Chongqing University Press*,pp.56-58.

[6]Wu Fubao, Li Qi, Song Wenzhong. A Approach of Inductive Learning of Expression system based on The Knowledge of the theory of rough set. *Control and Decision*, 1999,14(3):205～ 211.

[7]Wang Hongwei.The Research of Classification Algorithm Based on Decision Tree. *Beijing: Software Guide*, 2007, pp.81-91.

[8]Ji Guishu, Chen Peiling, Song Hang. General Description of The Research of Classification Algorithm of Decision Tree, *Technology Square, 2007*,pp.29-31.

[9]Zhao Weidong,Sheng Zhaohan. A Approach of Inductive Learning based on The theory of rough set. *Learned Journal of Management Engineering* ,2000,pp.289-301.

[10]HON G J R. A E1, an extension matrix approximate method for the general covering problem.*International J Computer and Information Sciences,* 1985, pp.89-101.